

Big Network Pipes for the ASCI Program

by Stephen Tenbrink, Deputy Group Leader, CIC-5 Network Engineering

The Accelerated Strategic Computing Initiative (ASCI) program, which has deployed new high-end computing systems at the three DOE Defense Programs (DP) laboratories, has another component requiring higher bandwidth between these laboratories. With the decision to place a 10-teraflop "ASCI White" computer at Livermore at the end of this year and a 30-teraflop computing system at Los Alamos at the end of 2001, there will be an imbalance of compute power among the three DP laboratories. The decision to create this imbalance was based on the need to avoid the high costs required to deploy these systems at all three sites. Along with this decision it was also apparent that the other two laboratories that did not have the current high-end ASCI system would need access to the machine and that this access should "appear as if local to the extent possible" to the remote user.

ASCI's Distance Computing

Out of this was born the "Distance and Distributed Computing" program also known as DisCom. The first phase of DisCom is to address the distance computing aspects of ASCI given the imbalance of computing power mentioned above. The second phase is to build upon the first phase by creating a wide area network (WAN)

computing fabric for the entire DOE Nuclear Weapons Complex that includes distributed computing resources located throughout the complex. Sandia National Laboratory is the lead lab for DisCom but both Los Alamos and Livermore have significant involvement in this effort.

So, what does it mean for a remote user to gain access that appears to the user "as if local..."? To respond to this requirement each lab interviewed some of their users to see what operations were needed to get meaningful results in a timely manner. (Note the words "timely manner" in the last sentence.) These operations involve data movement, as would be expected, but more important it's what is done with the data at the remote site that is critical. This could involve local archival storage, intermediate storage, high-resolution visualization with various forms of rendering, etc. After the interviews and comparing results from the other labs, the DisCom team found the answer was "all of the above." In other words, DisCom will need to create a WAN that interconnects the network backbones at each of the three laboratories (and an adjunct link to Sandia/California) with enough bandwidth that will support any of the network functions that users now do locally. To achieve this, very high-speed network data pipes would be needed. The only problem that DisCom will not be able to overcome is the latency issue due to the "time of flight" of the data from one site to another.

This could be tens of milliseconds if the user is in California and the ASCI computer is at Los Alamos. This is why the phrase "to the extent possible" was added to the DisCom goal.

How Big the Bandwidth?

DisCom project personnel then determined the size of the data pipes. This was done several ways but the guiding principal was to move "x" terabytes of data between sites in about an hour where "x" was the expected size of the problem data sets anticipated from the current and future ASCI high-end machines. While this is not an exact mathematical approach it would give us a good estimate of what bandwidth would be needed. The result was the approach showed that the minimum bandwidth should near the teraflop rating of the machine in gigabits/sec. Thus, for the 10-teraflop Livermore White computer, a 10-gigabit/sec link should suffice; for the Los Alamos 30-teraflop, a 30-gigabit/sec link; and for the future 100-teraflop ASCI machine, a 100-gigabit/sec link. When compared to the current interlab 155 megabits/sec WANlink (provided by ESNet) that LANL has, these rates seem astounding. But given the size of the data sets expected from the future ASCI machines and the goal of creating an "as if local" environment, these rates would be required.

The Telcos' Response to Future Growth

Luckily, the growth of the Internet is helping DisCom reach these bandwidth goals. Because of the high demand for digital communications for Internet activity, most major long distance telecommunication company carriers (Telcos) are installing new fiber optic links around the country and, as they install these links, they are adding extra capacity (addition fiber optic strands) for future growth. Accompanying this is a new fiber optic technology called Wave Division Multiplexing (WDM) and Dense Wave Division Multiplexing (DWDM) which allow one fiber strand to carry multiple data streams on different wavelengths (colors). These two efforts are combining to drive down the cost of wide area bandwidth to a point that even 100 gigabits/sec between California and New Mexico is within the budgetary constraints of DisCom. Two years ago it was thought that the Telcos would have trouble even providing 100 gigabit/sec. Today most of the Telcos we've approached hardly blink an eye when we state our requirements.

LANL, however has a unique problem in this area. Most long distance carriers have a "point of presence" (POP) in major metropolitan areas and rely on the local exchange carrier (LEC) to go the "last mile" from the POP to the final destination. Being that Los Alamos is where it is, this becomes a "last 100 mile" problem for the Laboratory. Our LEC, US West, provides the communication infrastructure for this last 100 miles that supports your long distance phone service and the Laboratory's ESNet link to a specific POP in Albuquerque depending on who the long distance carrier is. The problem is that we do not know if US West has the fiber capacity to support 30 or 100 gigabit/sec, even with DWDM. The answer to this question will be determined in the next few weeks when

responses to an Request For Quotation (RFQ) for the DisCom WAN are evaluated by DisCom TriLab members. Luckily, again because of the growing demand for Internet access, other companies are interested in building fiber links into northern New Mexico. These companies are also bidding on the DisCom RFQ. It is hoped that the result of all of this will be a WAN that provides the required DisCom bandwidth and, in addition, provides the infrastructure to support increased bandwidth into northern New Mexico to meet the growing demand for Internet access, both commercial and residential.

For more information about the TriLab DisCom project, see the DisCom Web site at <http://www.cs.sandia.gov/discom/>, or contact Steve Tenbrink at sct@lanl.gov.

